Title

Evidence for negative selection in human primary fibroblasts to tolerate high somatic mutation loads upon treatment with multiple low doses of N-ethyl-N-nitrosourea

Authors

Johanna Heid[1,=,*], Ronald Cutler[1,2,=,*], Shixiang Sun[1], Moonsook Lee[1], Alexander Y. Maslov[1,3], Xiao Dong[4,5], Simone Sidoli[2] and Jan Vijg[1,6,*]

Affiliations

[1] Department of Genetics, Albert Einstein College of Medicine, Bronx, NY 10461, USA

[2] Department of Biochemistry, Albert Einstein College of Medicine, Bronx, NY 10461, USA

[3] Laboratory of Applied Genomic Technologies, Voronezh State University of Engineering Technology, Voronezh 394000, Russia

[4] Institute on the Biology of Aging and Metabolism, University of Minnesota, Minneapolis, MN 55455, USA

[5] Department of Genetics, Cell Biology and Development, University of Minnesota, Minneapolis, MN 55455, USA

[6] Center for Single-Cell Omics, School of Public Health, Shanghai Jiao Tong University School of Medicine, Shanghai 200025, China

[=] Authors contributed equally to this work

[*] Correspondence to J.H. (johanna.heid@einsteinmed.edu), R.C. (ronald.cutler@einsteinmed.edu, @cutleraging), or J.V. (jan.vijg@einsteinmed.edu)

Summary

High-throughput sequencing at the single-cell and single-molecule level has shown that mutation rate is much higher in somatic cells than in the germline, with thousands of mutations accumulating with age in most human tissues. While there is now ample evidence that some of these mutations can clonally amplify and lead to disease, most notably cancer, the total burden of mutations a cell can tolerate without functional decline remains unknown. Here we addressed this question by exposing human primary fibroblasts multiple times to low doses of N-ethyl-N-nitrosourea (ENU) and quantitatively analyzing somatic mutation burden using single-cell whole genome sequencing. The results indicate that individual cells can sustain ~60,000 single-nucleotide variants (SNVs)

with only a slight adverse effect on growth rate. We found evidence for selection against potentially deleterious variants in gene coding regions as well as depletion of mutations in sequences associated with genetic pathways expressed in these human fibroblasts, most notably those relevant for maintaining basic cellular function and growth. However, no evidence of negative selection was found for variants in non-coding regions. We conclude that actively proliferating fibroblasts can tolerate very high levels of somatic mutations without major adverse effects on growth rate via negative selection against damaging coding mutations. Since most tissues in adult organisms have very limited capacity to select against mutations based on a growth disadvantage, these results suggest that a causal effect of somatic mutations in aging and disease cannot be ruled out.

Key words

Aging, Somatic mutations, Mutation burden, N-ethyl-N-nitrosourea, Single-cell whole-genome sequencing, Negative selection, Coding regions, Non-coding regions, Human lung fibroblasts

Introduction

Aging is a complex process of continuous decline in cellular and tissue function that is accompanied with increased disease risk (López-Otín, Blasco et al. 2023). DNA damage is thought to play a central role in this process and may be a universal cause of aging (Schumacher, Pothof et al. 2021). Among the major molecular consequences of DNA damage during aging is an increase in somatic DNA mutations. Somatic mutations are the result of errors during repair or replication of a damaged DNA template, and they cannot be corrected except by the death of the cell or organism. The accumulation of somatic mutations has since long been implicated as a major cause of aging (Failla 1958, Szilard 1959), which is strongly supported by the exponential increase of cancer with age (Laconi, Marongiu et al. 2020). However, until recently, cancer is the only age-related disease that has been conclusively demonstrated to be caused by somatic mutations. In part, this is due to the lack of reliable assays to quantitatively analyze somatic mutations in normal tissues. Recently, this situation has changed and single-cell and single-molecule assays have emerged that have now conclusively shown that mutations do indeed accumulate with age in most if not all human tissues (Ren, Dong et al. 2022), including postmitotic tissues such as the brain (Lodato, Rodin et al. 2018). Evidence has now also been provided for somatic mutations as the cause of a wide variety of human diseases other than cancer (Erickson 2010, Li and Williams 2013) (Mustjoki and Young 2021). Like cancer, many of these age-related effects may be due to clonally amplified mutations. Indeed, in blood, somatic mutations lead to clonal hematopoiesis,

which has been associated with increased risk of age-related diseases, such as cardiovascular disease, as well as with overall mortality (Jaiswal, Fontanillas et al. 2014).

What thus far remains unknown is the possible functional impact, collectively, of thousands of random mutations accumulating in a cell, for example, by adversely affecting gene-coding and gene-regulatory sequences relevant for the cell type studied. If so, one would expect a limit to the burden of random mutations a cell can tolerate. To test this, we treated IMR-90 human primary fibroblasts with low, repeated doses of N-ethyl-N-nitrosourea (ENU), a powerful mutagen that induces almost exclusively single-nucleotide variants (SNVs). Our results show that repeated ENU treatment linearly increased mutation frequency up to ~60,000 SNVs per cell, without a major effect on growth rate. We found evidence for selection against potentially deleterious mutations in gene coding regions and functional genomic regions associated with cell growth, but did not find evidence for this in non-coding regions. Therefore, this suggests that accumulating somatic mutations do have functional effects when negative selection is not possible, as is the case in most cells and tissues in adult organisms (Sender and Milo 2021).

Results

Repeated Mutagen Treatment Only Slightly Affects Cell Growth

To determine the impact of increased somatic mutation burden in primary human cells *in vitro*, we treated human fetal lung fibroblasts (IMR-90) a total of 9 times with a sublethal dose of 50 mg/ml of N-ethyl-N-nitrosourea (ENU), each treatment was followed by a recovery period (Fig. 1A; see also Methods for details). To monitor cell growth, toxicity, and mutation burden, we measured cell numbers and apoptosis rates at 3 and 7 days after treatment. The study design is schematically depicted in Fig. 1A. We found that repeated treatment with ENU slightly decreased cell growth rate measured at 3 days (Supplementary Fig. 1A; $p=3.32*10^{-2}$, 2-way ANOVA) and 7 days (Supplementary Figs. 1B; $p= 1.83*10^{-2}$, 2-way ANOVA) resulting in a total of 28.25±1.87 population doublings (PDs) for control cells and 24.1±1.34 PDs for ENU-treated cells. As expected, this effect was stronger closer to the treatment, i.e., at the 3-day time point (Supp.Fig.1A) compared to the 7-day time point (Supp.Fig1B). The reduced growth rate of the mutagen-treated cells as compared to the controls is likely due to a longer PD time at the 3-day timepoint (Supplementary Fig.1C) and the DNA damage response, as indicated by an increase in early- and late-stage apoptotic cells at both the 3 and 7 day time points, measured by Annexin and 7-amino-actinomycin D staining, respectively (Fig. 1C-F, Supplementary Fig. 1E-H; $p < 0.0001$ for Annexin and <0.0001 for 7-AAD; paired t-test). Evidence for the activation of a DNA damage response was also obtained by global RNA sequencing of ENU-treated cells at cycle 9 as compared to control cells. Gene set enrichment analysis using edgeR showed mismatch repair (MMR) and

nucleotide excision repair (NER) pathways enriched in the ENU-treated cells (Fig.1H, see also Supplementary Fig.2A-C for gene expression analysis of the three independent replicates).

Repeated Mutagen Treatment Increases Mutation Burden by more than 30-fold

Since somatic DNA mutations are random, they cannot be distinguished from sequencing artifacts after direct whole genome sequencing. Hence, we used our established and highly accurate single-cell whole genome sequencing assay and SCcaller 2.0 (Dong, Zhang et al. 2017) to analyze mutation burden in 21 single-cells from control and ENU-treated groups collected in 3 independent batches at cycles 3, 6, and 9 at the 7 day time point (Fig. 1A, Supplementary Table 1). Mutation frequencies were assessed at 7 days after each treatment, as DNA damage responses should be abated by that time and all mutations fixed. Following quality control, mutation filtering, and correction for genome coverage and detection sensitivity (Supplementary Figs. 3A-G, Methods), we found the number of estimated single nucleotide variants (SNVs) per cell in the control group increased from 2,434±406 at cycle 1 to 19,064±823 in the ENU-treated group at cycle 3, i.e. a 7.8-fold increase ($p=2.19 *10^{-6}$, two-sided linear model with estimated marginal means, Tukey's HSD) (Fig. 2A). At cycle 9, there was a 32-fold increase in estimated SNVs per cell in the ENU-treated group relative to the controls ($p=2.28*10^{-14}$, two-sided linear model with estimated marginal means, Tukey's HSD), where the total amount of estimated SNVs per cell in the ENU-treated group reached a maximum of 55,954±4,066 (Fig. 2A). The increase in SNVs over the course of the experiment is best fit by a linear model (Supplementary Fig. 4A). Importantly, there was no significant difference between control cells at cycle 1 and 9 (p=0.99, two-sided linear model with estimated marginal means, Tukey's HSD), ruling out that the increase in SNVs in ENU-treated cells was due to passaging (Fig. 2A). These results were orthogonally validated with our recently developed single-molecule assay (Maslov, Makhortov et al. 2022) (Supplementary Note 1, Supplementary Figs. 4B-E).

While ENU is known to mainly induce point mutations through alkylation damage, the estimated number of small insertions and deletions (INDELs) per cell was also observed to increase with ENU treatment, from 123±49 in the controls to 246±49 in the ENU-treated group at cycle 3, without further increase, i.e. a 1.8-fold increase ($p=2.74*10^{-4}$, two-way ANOVA) (Fig. 2B). While rare, INDELS are induced by ENU, albeit at low frequencies(Watson, Cunningham et al. 1998), and could be caused by DNA polymerase slippage during DNA replication at sites of alkylation damage. As with SNVs, there was no significant difference in the amount of estimated INDELs between control cells from cycle 1 and cycle 9 (p=0.79, two-sided linear model with estimated marginal means, Tukey's HSD) (Fig. 2B). However, as opposed to the linear increase in SNVs with increasing cycles, there was no further increase in the estimated INDELs per cell in the ENU-treated cells at cycle 6 or cycle 9 relative to cycle 3 (cycle 6 p=0.99; cycle 9 p=0.93, two-sided linear model with estimated marginal means, Tukey's HSD), suggestive of negative selection acting to prevent a further increase. This can also be seen by the downward trend of the

140 INDEL/SNV ratio in the ENU-treated cells from cycle 3 to cycle 9 (p=0.89, two-sided linear model
141 with estimated marginal means, Tukey's HSD) (Fig. 2C).

142 We then characterized the underlying SNV mutational processes by combining cells from all
143 cycles within the control and ENU-treated groups and found increases in the relative contribution
144 of thymine-to-adenine (T>A) and thymine-to-cytosine (T>C) base transversions (Fig. 2D,
145 Supplementary Fig. 5A). This is in keeping with previously published ENU-induced mutational
146 spectra by us and others (Barbaric, Wells et al. 2007, Maslov, Makhortov et al. 2022). We then
147 inferred SNV mutational signatures *de novo* using non-negative matrix factorization
148 (Supplementary Fig. 5B) and identified two signatures, named SBSA and SBSB, which were
149 dominant in the cells of the control and ENU-treated groups, respectively (Fig. 2E). SBSA
150 resembles COSMIC human signature SBS40 (cosine similarity=0.88, residual sum of
151 squares=$4.03*10^{-3}$) (Fig. 2F), while SBSB resembles COSMIC human signature SBS5 (cosine
152 similarity=0.71, residual sum of squares=$1.1*10^{-2}$) (Fig. 2G). Both mutational signatures are
153 known to correlate with aging (Alexandrov, Kim et al. 2020). Importantly, we also found that
154 SBSB resembled the SIGNAL human signature associated with ENU-treatment, validating the
155 source of the observed mutations (cosine similarity 0.86, residual sum of squares=$7.54*10^{-3}$) (Fig.
156 2H) (Kucab, Zou et al. 2019). Though relatively few INDELs were observed (369 total), we
157 provide a characterization of their mutational spectra and signature (Supplementary note 2,
158 Supplementary Fig. 2E).

159

160 Negative Selection Against Damaging Coding Variants

161 Based on our results showing that the mutation burden of primary cells could reach an estimated
162 ~60,000 SNVs per cell, i.e., ~10 times the SNV burden typically found in cells from aged tissues
163 (Ren, Dong et al. 2022), without a substantial effect on cell growth (Fig. 1B), we hypothesized
164 that this could be explained in part by negative selection. To test this, we annotated the 109,018
165 observed mutations amongst the 21 cells from both control and ENU-treated groups
166 (Supplementary Figs. 3H-I) into different types of potentially damaging coding region variants
167 (i.e., synonymous, missense, etc.) using the variant effect prediction program (McLaren, Gil et al.
168 2016). We then determined the frequency of these variants as well as the observed/expected (O/E)
169 ratios for each cell of the ENU-treated groups. The expected variant frequency was estimated by
170 recreating 1,000 instances of each cell where mutations are randomly distributed across the
171 genome to simulate the mutation distribution that would be expected under the null hypothesis,
172 i.e. neutral selection, while accounting for sequencing coverage, mutational signature, and
173 nucleotide context (Methods) (Bergstrom, Barnes et al. 2020).

174 We first assessed selection by calculating the ratio of non-synonymous variants to synonymous
175 variants (N/S). It should be noted that variants of this method, such as the dN/dS ratio, have been
176 used to identify selection in cancer studies (Martincorena, Raine et al. 2017). However, unlike data
177 on mutations coming from cancer clones, our data is limited in the frequency of coding mutations

178 and exhibits variable sequencing coverage across the genome, both of which are assumptions when
179 calculating the dN/dS ratio. This analysis revealed dose-dependent increases in frequency of both
180 synonymous ($p=1.27*10^{-6}$, two-way ANOVA) and non-synonymous variants ($p=1.11*10^{-5}$, two-
181 way ANOVA) in the ENU-treated group relative to the controls (Fig. 3A). However, there were
182 no significant differences in the N/S ratio either between the ENU-treated and control groups or
183 amongst the cycles within the ENU-treated group (Fig. 3B), suggesting that global selection
184 pressure for coding variants was unchanged in the context of very high mutation burden.

185 O/E ratios of coding region variants are particularly useful for assessing the presence of selection
186 only after accounting for the bias in mutation frequency due to DNA repair mechanisms, such as
187 transcription coupled repair, which is known to reduce mutation frequency in transcribed regions
188 and is agnostic to damage that would result in synonymous or non-synonymous variants. To
189 account for this, we normalized the O/E ratio of coding variants by the O/E ratio of intron
190 mutations, whose occurrence is dependent on DNA repair mechanisms, but not selective
191 mechanisms, as intronic mutations do not affect the downstream protein. This analysis revealed a
192 significant enrichment of non-synonymous variants at cycle 9 in the control group ($p=4.12*10^{-2}$,
193 Fisher's combined test statistic of permutation tests on individual cells, FDR) and a depletion of
194 non-synonymous variants at cycles 6 and 9 in the ENU-treated group (cycle 6 $p=8.38*10^{-3}$; cycle
195 9 $p=1.86*10^{-6}$, Fisher's combined test statistic of permutation tests on individual cells, FDR) (Fig.
196 3C), indicative of negative selection against these potentially damaging coding variants. These
197 O/E ratios at cycles 6 and 9 in the ENU-treated group correspond to a reduction of 15.14±7.84 and
198 20.25±13 non-synonymous variants per cell, respectively. This is further supported by a significant
199 decrease in the N/S O/E ratio at cycle 9 in the ENU-treated group ($p=1.12*10^{-2}$, Fisher's combined
200 test statistic of permutation tests on individual cells, FDR) and the insignificant deviation of the
201 O/E ratio from 0 for synonymous variants or N/S ratio in both control and ENU-treated groups
202 (Figs. 2C-D), which we would expect to not be subject to selective forces. Of the non-synonymous
203 variants, the pattern of negative selection in the ENU-treated group could mostly be attributed to
204 missense variants, although the same trend (although not significant owing to a limited number of
205 observed variants) could be seen for start lost, stop lost, but not for stop gained variants
206 (Supplementary Figs. 6A-B). Taken together, these results suggest that damaging variants in
207 coding regions do not accumulate to levels that would be expected in the context of the very high
208 ENU-induced mutation burden and instead are subject to negative selection. This may explain how
209 the cell population treated with ENU was able to avoid dramatic declines in growth rate even as
210 overall SNV burden continued to increase linearly throughout the experiment.

211 Given the finding of negative selection in the coding regions of the genome, we tested for negative
212 selection in the non-coding portion of the genome that contains the vast majority of all mutations
213 (99.03%) (Supplementary Fig. 8A). For this analysis, we annotated SNVs using scaled CADD
214 (Combined Annotation Dependent Depletion) scores, which is a measure that integrates multiple
215 annotations to quantify how deleterious or functionally impactful a genetic variant is, with higher
216 scores indicating variants predicted to be more damaging (Kircher, Witten et al. 2014). We then

217 divided variants into 10 bins based on their CADD scores and calculated the frequency and O/E
218 ratio of variants within each bin (Methods). To validate this approach, we examined the binned
219 CADD score O/E ratios for all coding variants in the ENU-treated group. Because we had
220 information on which variants were synonymous and non-synonymous, we calculated the median
221 CADD score for each of these respective variants and found non-synonymous variants to be
222 significantly higher than synonymous variants ($p=8.84*10^{-14}$, two-sided linear model with
223 estimated marginal means, Tukey's HSD), as would be expected (Supplementary Fig. 8B). Indeed,
224 the CADD score O/E ratio for coding variants at the bin corresponding to the median CADD score
225 for synonymous variants (7.7-12.8) was indicative of neutral selection in both control and ENU-
226 treated groups, while the O/E ratio for coding variants at the CADD score bin corresponding to
227 the non-synonymous variants (21.4-35.7) was indicative of negative selection at cycles 6 and 9 in
228 the ENU-treated group (cycle 6 $p=3.94*10^{-3}$; cycle 9 $p=6.29*10^{-3}$ Fisher's combined test statistic
229 of permutation tests on individual cells, FDR) (Supplementary Fig. 8C). Because this result shows
230 that variants with higher CADD scores are under negative selection in the ENU-treated group
231 (validating our previous results showing non-synonymous variants to be under negative selection),
232 we reasoned that this score could allow us to dissect the varying effects of non-coding variants and
233 then test if non-coding variants were under negative selection in the context of high mutation
234 burden in the ENU-treated group. Thus, we classified each non-coding SNV based on their
235 genomic position (e.g. 5KB upstream, regulatory, splice, non-coding RNA, etc.) and examined the
236 binned CADD score O/E ratios for patterns of negative selection in the ENU-treated group.
237 Interestingly, we did not find any evidence of negative selection among any of the classified non-
238 coding variants, with the exception of those in introns, which was expected due to transcription
239 coupled repair (Supplementary Fig. 8D). Taken together, these results reinforce the finding that
240 deleterious variants in coding regions are subject to negative selection, which delays their
241 accumulation in a population of proliferating cells, while non-coding variants are under neutral
242 selection which allows them to accumulate in non-coding regions that are known to play a role in
243 regulating gene expression. This is likely due to the relatively small effect of each individual non-
244 coding variant (Supplementary Fig. 8B). However, because the majority of SNVs accumulate in
245 non-coding regions (Supplementary Fig. 9A), this does not rule out a possible cumulative effect
246 that these variants may have on downstream gene regulation.

247

248 <u>Protection Against ENU-induced Mutation Accumulation in Functional Pathways</u>

249 To investigate if negative selection acts at the pathway level under high mutation burden, we
250 analyzed the distribution of SNVs across genomic regions functionally relevant for the primary
251 fibroblast model under study, i.e., pathways involved in cell growth and basic cellular functions.
252 For this purpose, we defined functional genomic regions, (exons, upstream promoter, and
253 enhancers) for each gene and then calculated the frequency and O/E ratios of SNVs for different
254 groupings of genes (Methods). We first reasoned that genes with sufficient expression would have
255 greater functional importance to the cell as compared to those that were not expressed, as has been

demonstrated previously (Schuster-Böckler and Lehner 2012). This was supported by O/E ratios indicating that SNV frequencies in functional regions associated with expressed genes were significantly lower than expected by chance at cycle 9 in the control group and all cycles in the ENU-treated group (cycle 3 $p=1.6*10^{-4}$; cycle 6 $p=3.6*10^{-6}$; cycle 9 $p=3.6*10^{-6}$, Fisher's combined test statistic of permutation tests on individual cells, FDR), while they occurred at the expected frequencies in functional regions associated with non-expressed genes in both groups (Fig. 4A). Furthermore, the O/E ratio of expressed genes determined to be 'essential' by genome-scale perturbation studies (Tsherniak et al., 2017) showed a significant depletion of SNVs in ENU-treated cells at cycle 9 ($p=2.49*10^{-4}$, Fisher's combined test statistic of permutation tests on individual cells, FDR) (Fig. 4B). These results indicate that negative selection against mutations in expressed genes and more specifically, essential pathways, is sufficient to endure high mutation burden.

Turning our attention to the pathway level, we identified 382 gene ontology biological processes to be significantly enriched (adjusted $p<0.05$) in the set of expressed genes from our RNA-seq data (Supplementary Figs. 2A-C) and then calculated the frequency and O/E ratios for each of these pathways. We focused the analysis on the ENU-treated cells due to limited observed mutations in the control group. This resulted in 24 pathways (ranging in size of 76 to 418 genes or $1.1*10^6$ to $7.1*10^6$ base pairs) that contained significantly less mutations than would be expected by random chance ($p< 0.05$, Fisher's combined test statistic of permutation tests on individual cells, FDR) (Supplementary Fig. 9A). As a negative control, we also calculated O/E ratios after randomly shuffling genes amongst the identified pathways, which served to filter out pathways with non-specific functional genomic regions (Supplementary Figs. 9B-C). Amongst the pathways with significant depletion of SNVs, relatively few genes were shared, indicating negative selection was occurring at distinct genomic regions (Supplementary Fig. 9D). Moreover, these pathways were characterized as those related to basic cellular processes supporting fibroblast cell growth, such as regulation of response to oxidative stress (GO:1902882), regulation of DNA replication (GO:0006275), actin filament organization (GO:0007015), and regulation of DNA repair (GO:0006282) (Figs. 4C). It is interesting to note that while we did detect that SNVs in these pathways were being acted on by negative selection, a significant increase in the frequency of SNVs for each of these pathways in the ENU-treated cells relative to the controls (54±18 SNVs) could still be detected (Supplementary Fig. 9E). In summary, these results suggest that in the context of very high mutations burden, pathways supporting cell proliferation and survival in culture are under a greater degree of negative selection, which highlights not only a mechanism in which dividing cells are able to survive in the context of high mutation burden, but also the impact that random somatic SNVs have on the functioning of genetic pathways, evidenced by their reduced accumulation.

Discussion

Somatic mutations are known since the 1970s to be the main cause of cancer{Nowell, 1976 #3518}. Mainly because of the exponential increase in cancer risk with age, aging itself has been suggested to be caused by the accumulation of somatic mutations{Failla, 1958 #59}{Burnet, 1974 #3412}. Testing this hypothesis has long been constrained by a lack of methods to quantitatively analyze somatic mutations in tissues and organs during aging. However, both single-cell and single-molecule methods have recently been developed to accurately quantify SNVs and INDELS. With this, we and others have demonstrated that somatic mutation burden is far higher than suspected based on the known germline mutation frequency and increases with age up to several thousands of mutations per cell in aged subjects depending on the tissue{Ren, 2022 #3935}. However, a causal relationship of increased mutation burden and age-related functional decline and disease risk remains untested. Instead, attention has been focused on clonally amplified somatic mutations and their causal relationship to non-cancerous and age-related diseases {Ng, 2021 #3972}.

In the present study we tested for a possible limit to somatic mutation accumulation in a relatively simple cell culture model in which high levels of somatic mutations were induced by multiple low doses of the powerful mutagen ENU. In this system we were able to linearly increase the somatic mutation burden to ~60,000 SNVs per cell, over 30-fold higher than untreated control cells. Remarkably, this supraphysiological mutation load only slightly impacted cell growth rate, suggesting that most cells can sustain an astounding level of genetic damage. The most straightforward explanation for these findings is that the mammalian genome is highly robust and that even very high numbers of point mutations are unlikely to adversely affect cellular function. Indeed, this would be supported by results from a recent study which observed as many as ~30,000 SNVs in intestinal crypts of patients with genetic defects in DNA mismatch repair {Robinson, 2021 #3973}.

An alternative explanation for the high tolerance of actively proliferating fibroblasts is negative selection against mutations affecting growth rate and cell viability. Indeed, we found multiple lines of evidence for negative selection acting on potentially deleterious coding region variants. While the frequency of both synonymous and non-synonymous SNVs increased linearly with ENU treatment, deeper analysis using observed/expected (O/E) ratios revealed a depletion of high impact non-synonymous variants, especially missense mutations, than would be expected by random chance in the ENU-treated cells. This suggests that actively proliferating fibroblasts have the capacity to selectively purge the most harmful coding mutations, preserving cellular fitness despite the vast number of random mutations accumulating genome-wide.

In contrast, no such negative selection was observed for SNVs in non-coding genomic regions, even when these variants were stratified by their predicted deleteriousness based on scaled CADD scores. This differential selection between coding and non-coding variants likely reflects the relatively large deleterious effect that coding mutations can have on protein function, compared to the typically smaller impact of non-coding variants on gene regulation. Nevertheless, the lack of selection against the vast majority of non-coding mutations does not

rule out the possibility of cumulative effects on cellular phenotypes over time, as has been postulated in the context of aging and cancer.

One approach to measuring the cumulative effects of mutations is to analyze their frequency at the pathway level, which revealed that functional genomic regions associated with genes involved in core cellular pathways, such as DNA replication, DNA repair, and cytoskeletal organization, showed a significant depletion of mutations in the ENU-treated cells than would be expected by random chance. This suggests that negative selection acts to preferentially maintain the integrity of pathways critical for cellular function and proliferation, despite the massive somatic mutational burden. Importantly, while the frequency of mutations in these pathways was significantly lower than expected, a substantial increase in mutations was still observed compared to controls, highlighting the potential for accumulated mutations to impact cellular fitness, even if purged below a critical threshold.

In summary, the present findings demonstrate that actively proliferating human cells possess a remarkable capacity to tolerate extremely high levels of random somatic mutations, primarily through selective negative selection against the most damaging variants in coding regions and preservation of genomic integrity in critical cellular pathways. These results have important implications for our understanding of the role of somatic mutations in aging and age-related diseases, where the ability to effectively select against deleterious mutations may be limited compared to the *in vitro* culture conditions examined here.

Data Availability

Data will be made available upon publication.

Acknowledgement

This study was supported by NIH grants: T32AG023475 to R.C., U19AG056278, U01HL145560, U01ES029519, P01AG017242, P01AG047200, RF1AG068908, and P30AG038072 to J.V. , the US Department of Defense (BC180689P1) grant to J.V., and the Michael J. Fox Foundation and The Paul F. Glenn Center for the Biology of Human Aging at Albert Einstein College of Medicine.

Author Contributions

A.Y.M., and J.V. conceived this study and designed the experiments. J.H., M.S., and A.Y.M performed the experiments. J.H., R.C., and S.S. analyzed the data. X.D. provided insights on the data analysis. J.H., R.C., S.S., and J.V. wrote the manuscript.

Declaration of interests

M.L., A.Y.M., X.D., and J.V. are co-founders and shareholders of SingulOmics Corp. A.Y.M., and J.V. are co-founders of MutagenTech Crop. Others declare no conflict of interest.

Declaration of generative AI and AI-assisted technologies in the writing process

The authors employed ChatGPT for spelling and grammar checks during this work's preparation. They have reviewed and edited the content accordingly, assuming full responsibility for the publication's content.

Material and methods

<u>Cell culture</u>

Human IMR90 obtained from the American Tissue Culture Collection (ATCC) were culture in DMEM (Dulbecco's Modified Eagle Medium, from ATCC) supplemented with 10% fetal bovine serum (FBS, Gibco/Life Technologies) and 1% Penicillin-Streptomycin solution (Gibco/Life Technologies) at 3% oxygen and 37°C. Cells were treated for a total of nine times over the course of several months with ENU. Here, cells were treated with medium containing 50mg/ml ENU 24 hours after seeding 1 million cells in a 10cm dish. After 3 days cells were counted, 1 million cells re-seeded and leftovers collected and stored at -80°C. The same procedure was repeated 7 days

384 after treatment. All downstream analysis was performed on samples collected 7 days after the last
385 treatment unless stated otherwise. All cell culture experiments were conducted in three
386 independent replicates.

387

388 Library preparations

389 DNA and RNA were extracted from the same cell pellets using the Zymo DNA/RNA extraction
390 kit (Microprep plus D7005). 200ng gDNA was used as input for library construction for SNV
391 analysis. Library preparations were performed as described elsewhere (Maslov, Makhortov et al.
392 2022). mRNA libraries were prepared and sequenced by Novogene. All sequencing was performed
393 by Novogene Corp. CA.

394

395 Single-cell isolation and library preparation

396 Single cells were isolated using the CellRaft AIR system (Cell Microsystems). The CellRaft array
397 was prepared following the manufacturer's instruction and washed three times for 3 minutes with
398 warm PBS before addition of cell suspension. 1000 cells were seeded in 1ml for one CellRaft
399 array. Cells were allowed to settle for 3-4 hours at 37°C before isolation. Before isolation the raft
400 was carefully washed with warm medium to remove potential debris and dead cells. Individual
401 rafts containing one cell were deposited in PCR tubes with 2.5ul PBS and stored at -80°C until
402 further processing.

403 Libraries for whole genome sequencing (WGS) were prepared according to the previously
404 described scMDA protocol using a NEBNext Ultra II FS kit (NEB). For quality control purposes,
405 a locus drop out test (LDO) was performed using primers and SYBR reagent (Gundry, Li et al.
406 2012). Library size was assessed with a TapeStation (Agilent) using HS5000 screen tape for
407 quality control of cDNA and HS1000 screen tape for all final libraries. All RNA-Seq and WGS
408 libraries were sequenced on an Illumina HiSeq platform by Novogene Corp Inc., CA.

409

410 Alignment for WGS

411 Sequencing reads were aligned to the human reference genome as previously described18. In
412 brief, paired-end sequencing reads were trimmed to remove adapter and low-quality reads by
413 Trim Galore (version 0.4.1) and then mapped to the human reference genome (GRCh37,
414 including decoy contigs) using BWA(mem; version 0.7.10) (ref. 52). PCR duplication were
415 marked using Samtools (version 0.1.19) (ref. 53). Realignment of reads and recalibrations of
416 base quality scores were performed using Genome Analysis Toolkit (GATK; version 3.5.0)
417 (refs. 18,54).

418

419 Germline variant analysis

420 Germline variants were identified as those with more than four variant-supporting reads at the
421 depth ≥5× and GATK quality score ≥30. Filtering for germline SNVs (qual by depth (QD) <2.0,
422 Fisher strand (FS) >60.0, RMS mapping quality (MQ) <40.0, MQRankSum <−12.5,
423 ReadPosRankSum <−8.0, strand odds ratio (SOR) >3.0) and INDELs (QD <2.0, FS >200.0,
424 ReadPosRankSum <−20.0, SOR >10.0) was set as recommended by GATK.

425

426 Calling somatic SNVs and INDELs in scWGS

427 Somatic mutations were called using SCcaller (version
428 1.2; https://github.com/biosinodx/SCcaller)46. The identification of mutation was based on
429 presence of at least four reads containing the mutation in genome regions with at least20× and
430 30× sequencing depth for SNVs and INDELs, respectively. Potential amplification artifacts were
431 filtered out based on assessment of allelic bias determined by heterozygous SNPs on autosomes
432 identified as described elsewhere20. A variant quality score ≥25 was chosen as a minimal for
433 INDEL calling.

434

435 Removal of technical artifacts in scWGS

436 Technical artifacts inherent to single-cell MDA library preparation steps, such as overheating
437 during cell lysis or whole genome amplification, can be identified by their specific mutational
438 signatures scE and scF (Petljak, Alexandrov et al. 2019). We checked for the presence of these
439 signatures and found scF to be present in control cells from cycle 1 (Supplementary Fig. 5). This
440 artifact is characterized by caused higher than expected cytosine-to-thymine (C>T) transversions
441 at GCN contexts (Supplementary Fig. 5). Thus, control cells from cycle 1 were excluded from
442 further downstream mutational signature analyses.

443

444 Estimating mutation frequencies in scWGS

445 The frequency of somatic SNVs/INDELs per nucleus was estimated with the normalization of
446 genomic coverage and sensitivity

447 $$\text{frequency of somatic SNVs|INDELs per nucleus} = \frac{\text{frequency somatic SNVs|INDELs}}{\frac{surveyed\ genome\ (bp)}{size\ of\ genome\ (bp)}}$$

448 The surveyed genome per nucleus was calculated as the number of nucleotides with mapping
449 quality ≥40 and coverage depth ≥20× for SNVs and ≥30× for INDELs, which are also found in
450 bulk. The sensitivity of de novo mutation calling in the single nuclei was estimated as the ratio of
451 the number of heterozygous SNPs detected in single nuclei to the total number of heterozygous

452 SNPs detected in bulk DNA at a minimum sequencing depth of 20× in bulk and 20× and 30× in
453 single-cells for SNV and INDEL, respectively (Supplementary Table). For INDELs, we also
454 required quality score of SNPs in single cells ≥25.

455

456 **References**

457 Alexandrov, L. B., J. Kim, N. J. Haradhvala, M. N. Huang, A. W. Tian Ng, Y. Wu, A. Boot, K. R.
458 Covington, D. A. Gordenin, E. N. Bergstrom, S. M. A. Islam, N. Lopez-Bigas, L. J. Klimczak, J.
459 R. McPherson, S. Morganella, R. Sabarinathan, D. A. Wheeler, V. Mustonen, G. Getz, S. G. Rozen
460 and M. R. Stratton (2020). "The repertoire of mutational signatures in human cancer." Nature
461 **578**(7793): 94-101.
462 Barbaric, I., S. Wells, A. Russ and T. N. Dear (2007). "Spectrum of ENU-induced mutations in
463 phenotype-driven and gene-driven screens in the mouse." Environ Mol Mutagen **48**(2): 124-142.
464 Bergstrom, E. N., M. Barnes, I. Martincorena and L. B. Alexandrov (2020). "Generating realistic
465 null hypothesis of cancer mutational landscapes using SigProfilerSimulator." BMC Bioinformatics
466 **21**(1): 438.
467 Dong, X., L. Zhang, B. Milholland, M. Lee, A. Y. Maslov, T. Wang and J. Vijg (2017). "Accurate
468 identification of single-nucleotide variants in whole-genome-amplified single cells." Nat Methods
469 **14**(5): 491-493.
470 Erickson, R. P. (2010). "Somatic gene mutation and human disease other than cancer: an update."
471 Mutat Res **705**(2): 96-106.
472 Failla, G. (1958). "The aging process and cancerogenesis." Ann N Y Acad Sci **71**(6): 1124-1140.
473 Gundry, M., W. Li, S. B. Maqbool and J. Vijg (2012). "Direct, genome-wide assessment of DNA
474 mutations in single cells." Nucleic Acids Res **40**(5): 2032-2040.
475 Jaiswal, S., P. Fontanillas, J. Flannick, A. Manning, P. V. Grauman, B. G. Mar, R. C. Lindsley, C.
476 H. Mermel, N. Burtt, A. Chavez, J. M. Higgins, V. Moltchanov, F. C. Kuo, M. J. Kluk, B.
477 Henderson, L. Kinnunen, H. A. Koistinen, C. Ladenvall, G. Getz, A. Correa, B. F. Banahan, S.
478 Gabriel, S. Kathiresan, H. M. Stringham, M. I. McCarthy, M. Boehnke, J. Tuomilehto, C. Haiman,
479 L. Groop, G. Atzmon, J. G. Wilson, D. Neuberg, D. Altshuler and B. L. Ebert (2014). "Age-related
480 clonal hematopoiesis associated with adverse outcomes." N Engl J Med **371**(26): 2488-2498.
481 Kircher, M., D. M. Witten, P. Jain, B. J. O'Roak, G. M. Cooper and J. Shendure (2014). "A general
482 framework for estimating the relative pathogenicity of human genetic variants." Nat Genet **46**(3):
483 310-315.
484 Kucab, J. E., X. Zou, S. Morganella, M. Joel, A. S. Nanda, E. Nagy, C. Gomez, A. Degasperi, R.
485 Harris, S. P. Jackson, V. M. Arlt, D. H. Phillips and S. Nik-Zainal (2019). "A Compendium of
486 Mutational Signatures of Environmental Agents." Cell **177**(4): 821-836.e816.
487 Laconi, E., F. Marongiu and J. DeGregori (2020). "Cancer as a disease of old age: changing
488 mutational and microenvironmental landscapes." Br J Cancer **122**(7): 943-952.
489 Li, C. and S. M. Williams (2013). "Human Somatic Variation: It's Not Just for Cancer Anymore."
490 Current Genetic Medicine Reports **1**(4): 212-218.
491 Lodato, M. A., R. E. Rodin, C. L. Bohrson, M. E. Coulter, A. R. Barton, M. Kwon, M. A. Sherman,
492 C. M. Vitzthum, L. J. Luquette, C. N. Yandava, P. Yang, T. W. Chittenden, N. E. Hatem, S. C.
493 Ryu, M. B. Woodworth, P. J. Park and C. A. Walsh (2018). "Aging and neurodegeneration are
494 associated with increased mutations in single human neurons." Science **359**(6375): 555-559.
495 López-Otín, C., M. A. Blasco, L. Partridge, M. Serrano and G. Kroemer (2023). "Hallmarks of
496 aging: An expanding universe." Cell **186**(2): 243-278.

Martincorena, I., K. M. Raine, M. Gerstung, K. J. Dawson, K. Haase, P. Van Loo, H. Davies, M. R. Stratton and P. J. Campbell (2017). "Universal Patterns of Selection in Cancer and Somatic Tissues." Cell **171**(5): 1029-1041.e1021.

Maslov, A. Y., S. Makhortov, S. Sun, J. Heid, X. Dong, M. Lee and J. Vijg (2022). "Single-molecule, quantitative detection of low-abundance somatic mutations by high-throughput sequencing." Sci Adv **8**(14): eabm3259.

McLaren, W., L. Gil, S. E. Hunt, H. S. Riat, G. R. Ritchie, A. Thormann, P. Flicek and F. Cunningham (2016). "The Ensembl Variant Effect Predictor." Genome Biol **17**(1): 122.

Mustjoki, S. and N. S. Young (2021). "Somatic Mutations in "Benign" Disease." N Engl J Med **384**(21): 2039-2052.

Petljak, M., L. B. Alexandrov, J. S. Brammeld, S. Price, D. C. Wedge, S. Grossmann, K. J. Dawson, Y. S. Ju, F. Iorio, J. M. C. Tubio, C. C. Koh, I. Georgakopoulos-Soares, B. Rodríguez-Martín, B. Otlu, S. O'Meara, A. P. Butler, A. Menzies, S. G. Bhosle, K. Raine, D. R. Jones, J. W. Teague, K. Beal, C. Latimer, L. O'Neill, J. Zamora, E. Anderson, N. Patel, M. Maddison, B. L. Ng, J. Graham, M. J. Garnett, U. McDermott, S. Nik-Zainal, P. J. Campbell and M. R. Stratton (2019). "Characterizing Mutational Signatures in Human Cancer Cell Lines Reveals Episodic APOBEC Mutagenesis." Cell **176**(6): 1282-1294.e1220.

Ren, P., X. Dong and J. Vijg (2022). "Age-related somatic mutation burden in human tissues." Front Aging **3**: 1018119.

Schumacher, B., J. Pothof, J. Vijg and J. H. J. Hoeijmakers (2021). "The central role of DNA damage in the ageing process." Nature **592**(7856): 695-703.

Schuster-Böckler, B. and B. Lehner (2012). "Chromatin organization is a major influence on regional mutation rates in human cancer cells." Nature **488**(7412): 504-507.

Sender, R. and R. Milo (2021). "The distribution of cellular turnover in the human body." Nat Med **27**(1): 45-48.

Szilard, L. (1959). "ON THE NATURE OF THE AGING PROCESS." Proc Natl Acad Sci U S A **45**(1): 30-45.

Tsherniak, A., F. Vazquez, P. G. Montgomery, B. A. Weir, G. Kryukov, G. S. Cowley, S. Gill, W. F. Harrington, S. Pantel, J. M. Krill-Burger, R. M. Meyers, L. Ali, A. Goodale, Y. Lee, G. Jiang, J. Hsiao, W. F. J. Gerath, S. Howell, E. Merkel, M. Ghandi, L. A. Garraway, D. E. Root, T. R. Golub, J. S. Boehm and W. C. Hahn (2017). "Defining a Cancer Dependency Map." Cell **170**(3): 564-576.e516.

Watson, D. E., M. L. Cunningham and K. R. Tindall (1998). "Spontaneous and ENU-induced mutation spectra at the cII locus in Big Blue Rat2 embryonic fibroblasts." Mutagenesis **13**(5): 487-497.

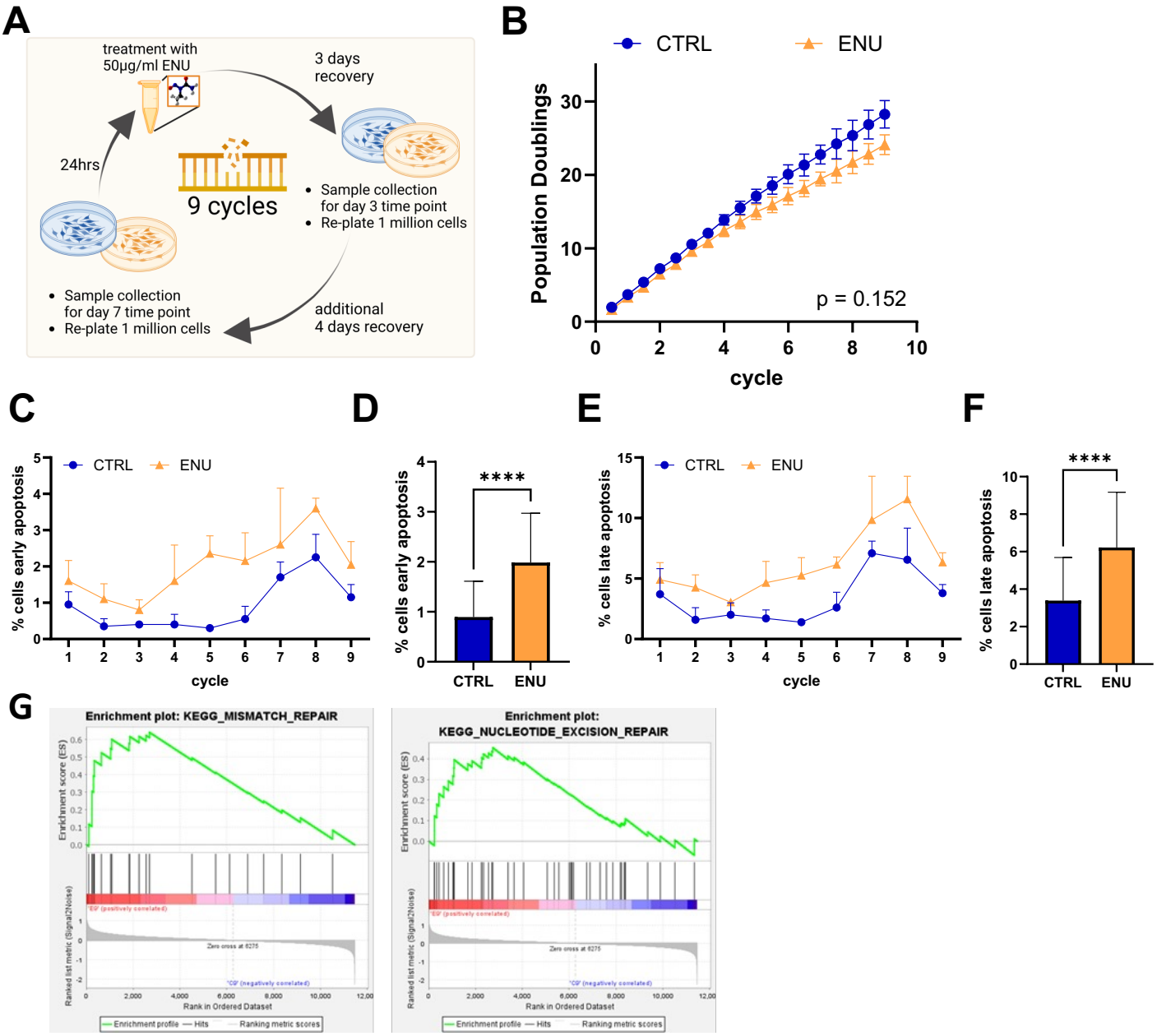


Figure 1. Repeated Mutagen Treatment Only Slightly Affects Cell Growth. A) To start the experiment, 1 million human primary lung fibroblasts (IMR90) cells were are plated and allowed to expand for 24 hours. Cells were then treated with a low dose of N-ethyl-N-nitrosourea (ENU) and allowed to recover for 72 hours (3 days). Following this cells were collected and assayed for cell number and cell viability. 1 million cells were re-plated and then allowed another 96 hours to recover and expand when they are then collected 7 days after treatment to again assess cell number and viability. Cell pellets were collected for transcriptome changes and somatic mutation burden. This procedure was repeated 9 consecutive times over the course of ~10 weeks. B) Combined cell growth data from early (3 days post-treatment) and late (7 days post-treatment) recovery time points showing total number of population doublings throughout entire experiment (2-way ANOVA to determine significance). See supplementary figure 1 for the separate assessments of the 3-day and 7-day recovery period (significance was tested with a 2-way ANOVA). C-E) Assessment of early (C-D) and late (E-F) apoptosis reveals contribution of increased cell death to slowed cell growth rate when measured at the 7-day time point. E and G) Average percentage of apoptotic cells across cycles in control (CTRL) and ENU cells. (N=3, mean and standard deviation; paired t-test was used to determine significance G) Gene set enrichment analysis (GSEA) results of samples treated with ENU show the enrichment of genes involved in DNA repair. P-value legend: **: p <= 0.005; ****: p<= 0.0001.

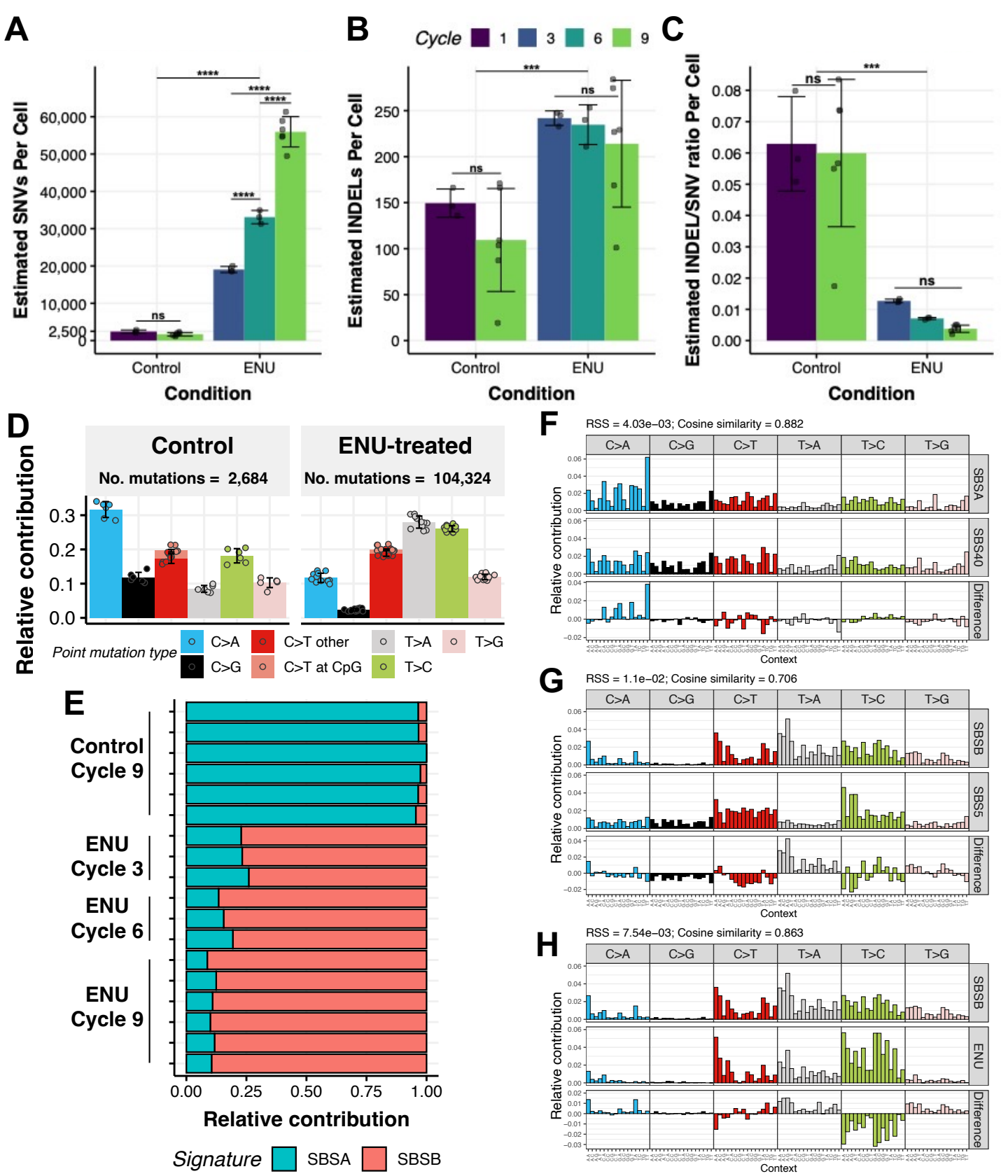


Figure 2. Single-cell somatic mutations analysis of repeated low-dose ENU treatment. A) Estimated burden of somatic single nucleotide variants (SNVs) and B) insertions/deletions and (INDELs) corrected for genome coverage and sensitivity. C) Ratio of INDELs to SNVs. D) Mutation spectrum of the relative contribution of SNV types grouped by condition. E) Relative contribution of *de novo* mutational signatures SBSA and SBSB to individual cells grouped by condition and cycle number. F-H) Comparison of *de novo* mutational signatures to most similar fitted signature from COSMIC (F-G) and SIGNAL (H) databases. Each point represents an individual cell and error bars represent 1 standard deviation. P-values comparing control and ENU-treated groups are shown at the top of plots A-C and estimated using a two-way ANOVA. Symbols at top of plots A-C indicate resulting p-values using a two-sided linear model with estimated marginal means and Tukey's HSD. P-value legend: ns: $p > 0.05$; ***: $p <= 0.001$; ****: $p <= 0.0001$

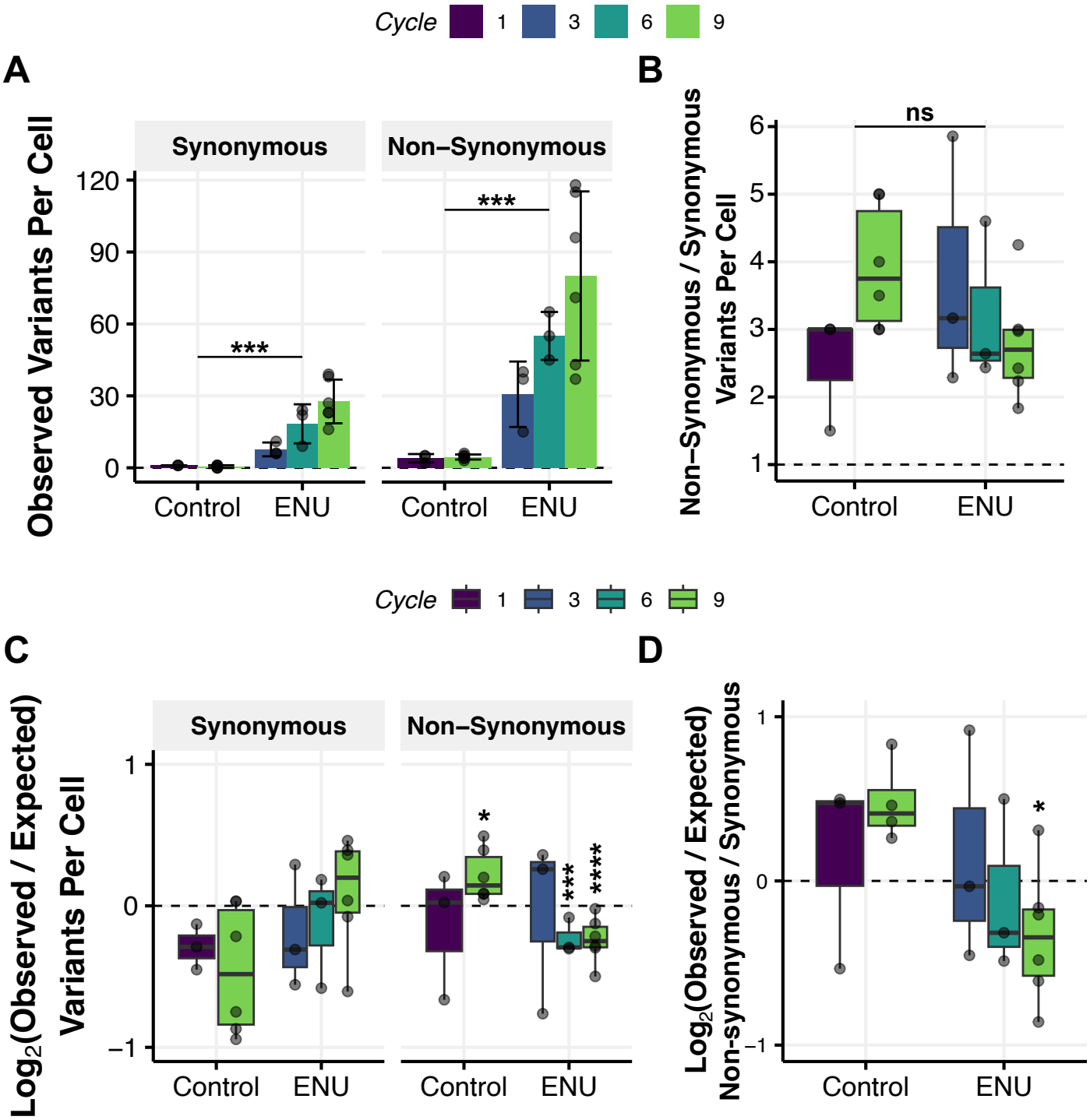


Figure 3. Negative Selection Against Damaging Coding Variants. A) Somatic mutation burden of synonymous and non-synonymous coding variants. B) Ratio of non-synonymous to synonymous variants. Symbols at top of plots A-B indicate resulting p-values of Mann-Whitney test comparing control and ENU-treated groups with multiple hypothesis correction. C) Log_2 observed vs expected of synonymous and non-synonymous coding variants. D)) Log_2 observed vs expected non-synonymous to synonymous variants. Symbols at top of plots C-D indicate resulting p-values of permutation test (N = 1000) comparing observed to expected. P-value legend: *: p <= 0.05; ***: p <= 0.001; ****: p <= 0.0001

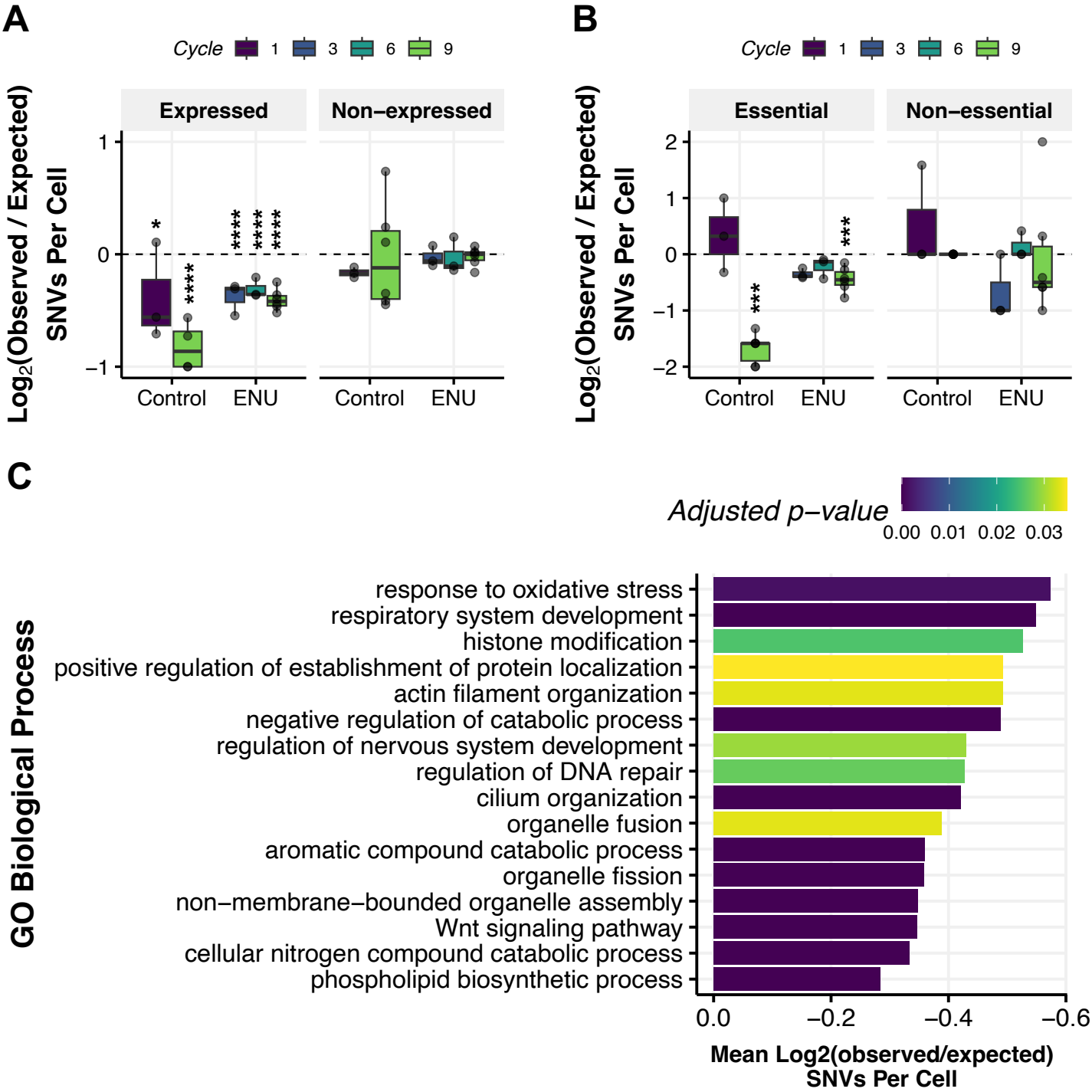
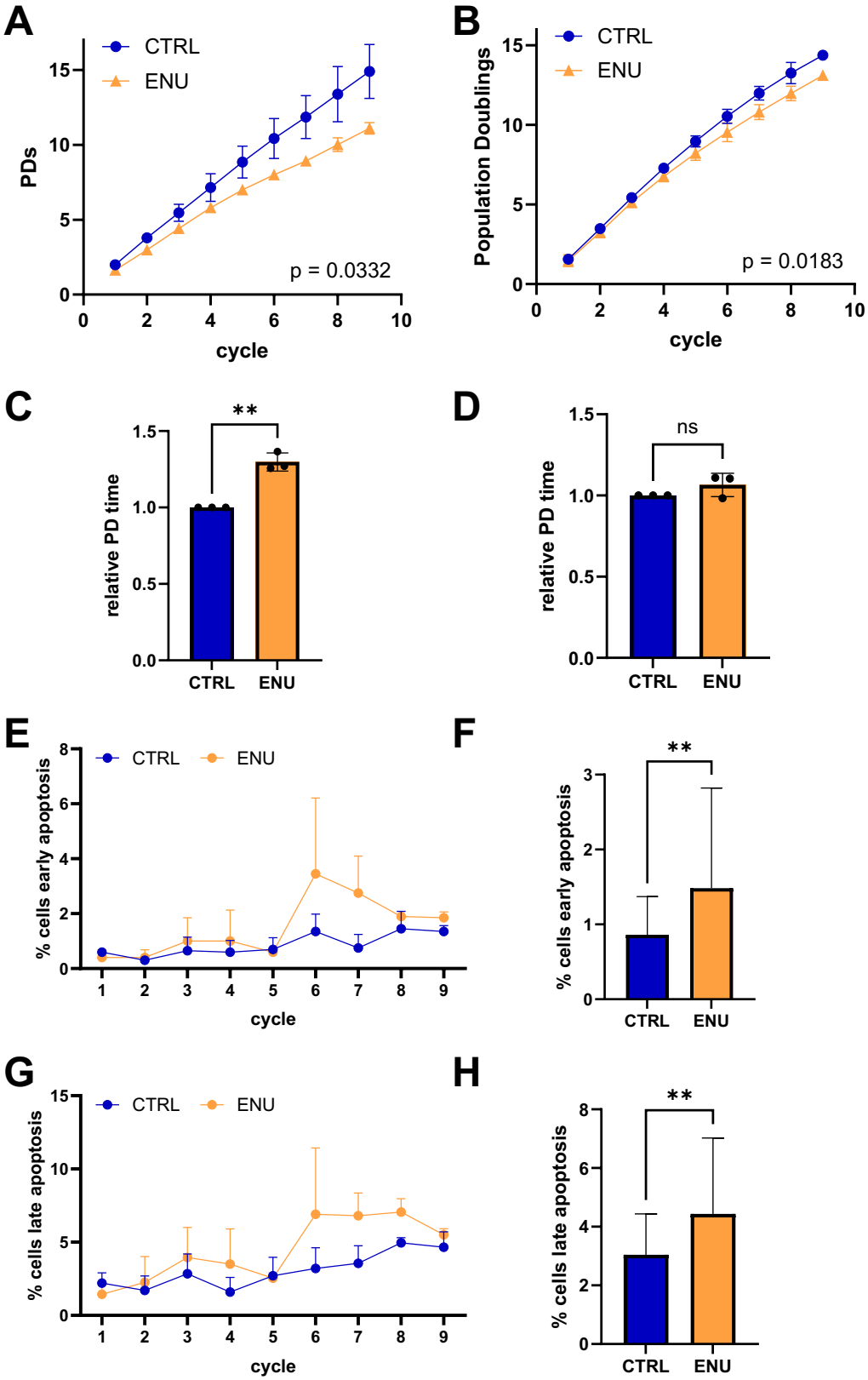
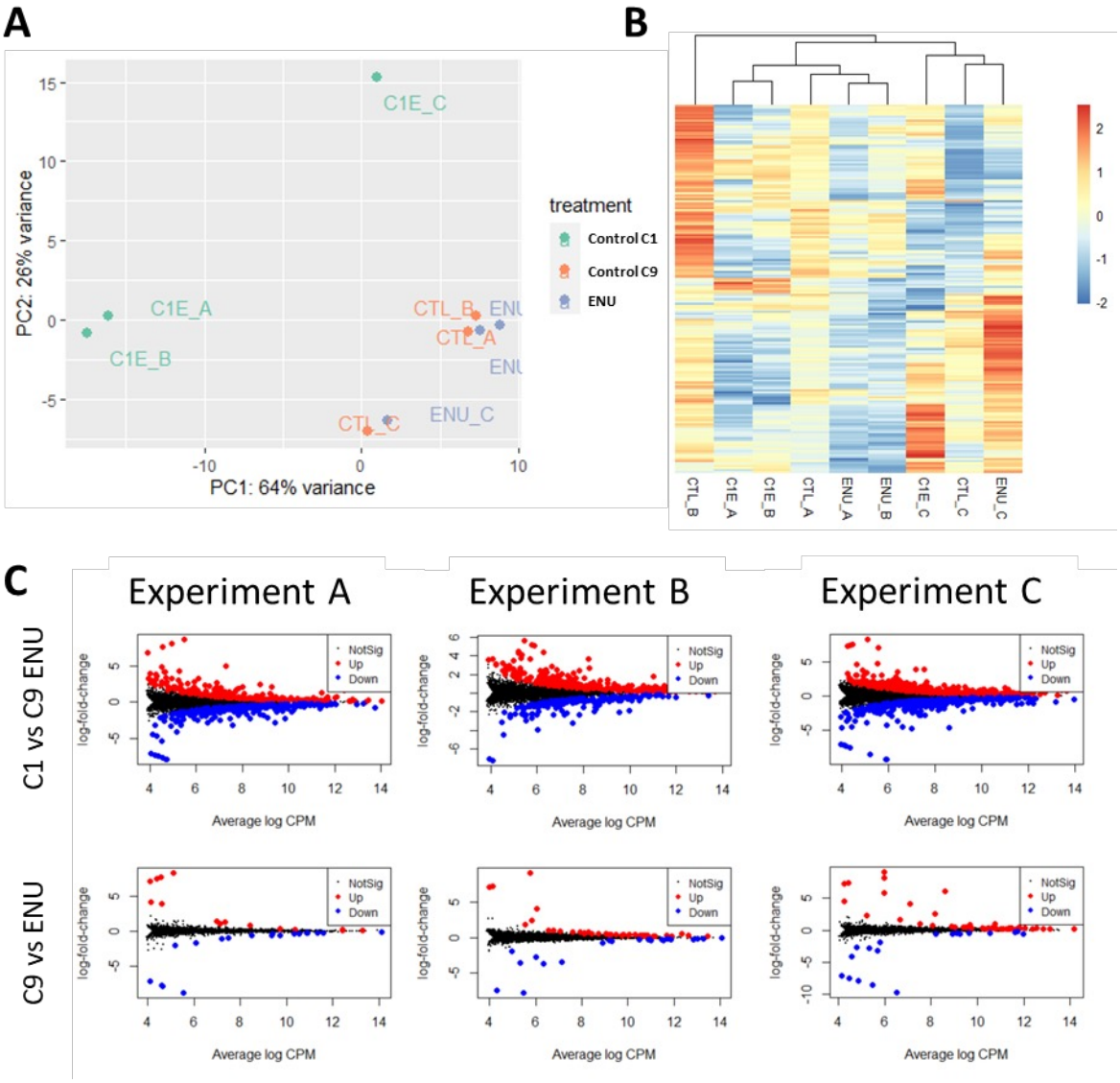


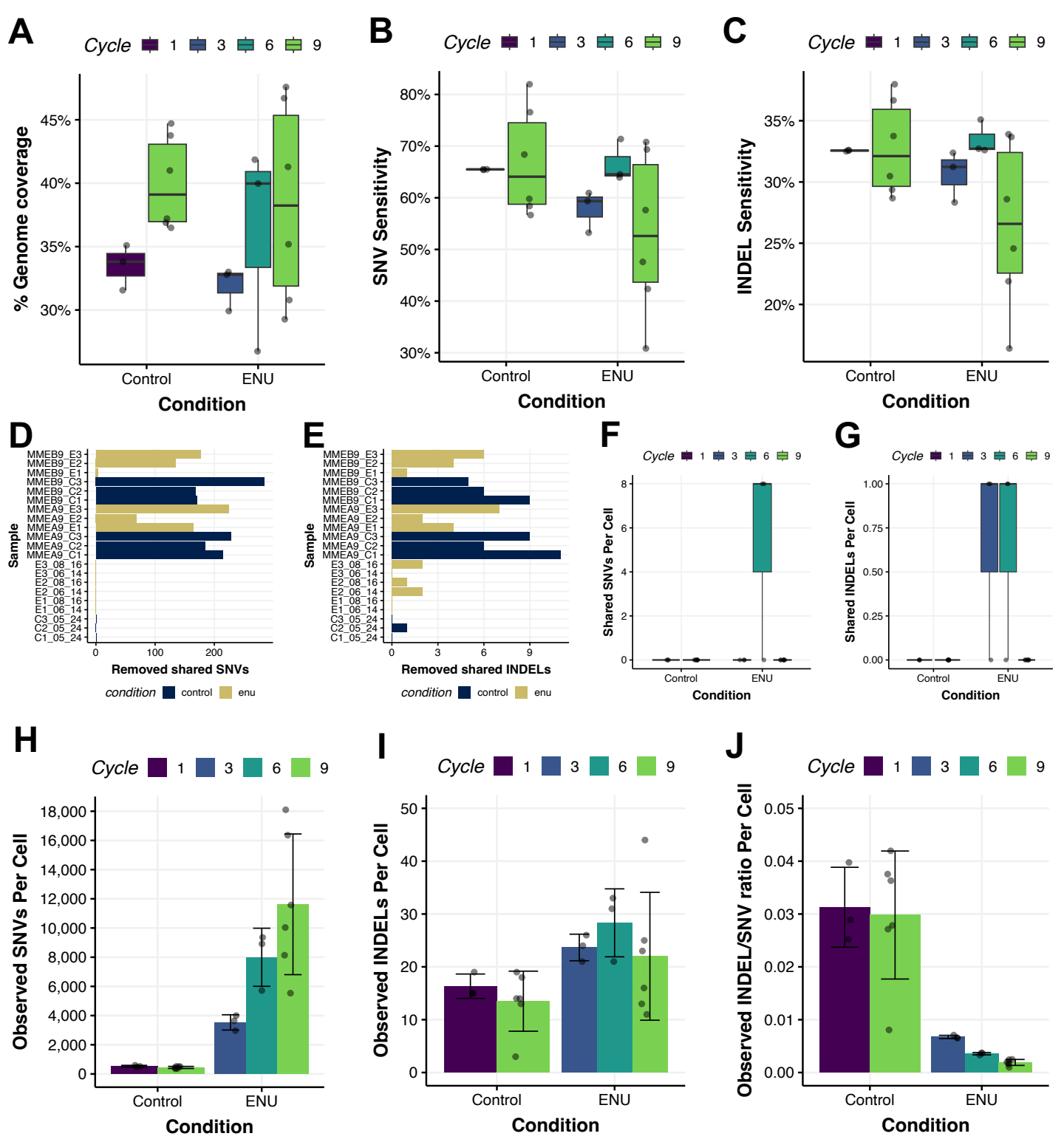
Figure 4. Protection Against ENU-induced Mutation Accumulation in Functional Pathways. A) Log_2 observed vs. expected ratios of SNVs in expressed or non-expressed genes in human lung fibroblasts. B Log_2 observed vs. expected ratios of SNVs in essential or non-essential genes expressed in human lung fibroblasts. Symbols at top of plots A-B indicate resulting p-values of permutation test (N = 1000) comparing observed to expected; C) Top GO biological processes in ENU-treated cells with significant depletion of SNVs indicated by Log_2 observed vs. expected ratios of SNVs in functional regions of each pathway. P-value legend: *: p <= 0.05; **: p <= 0.01; ***: p <= 0.001; ****: p <= 0.0001



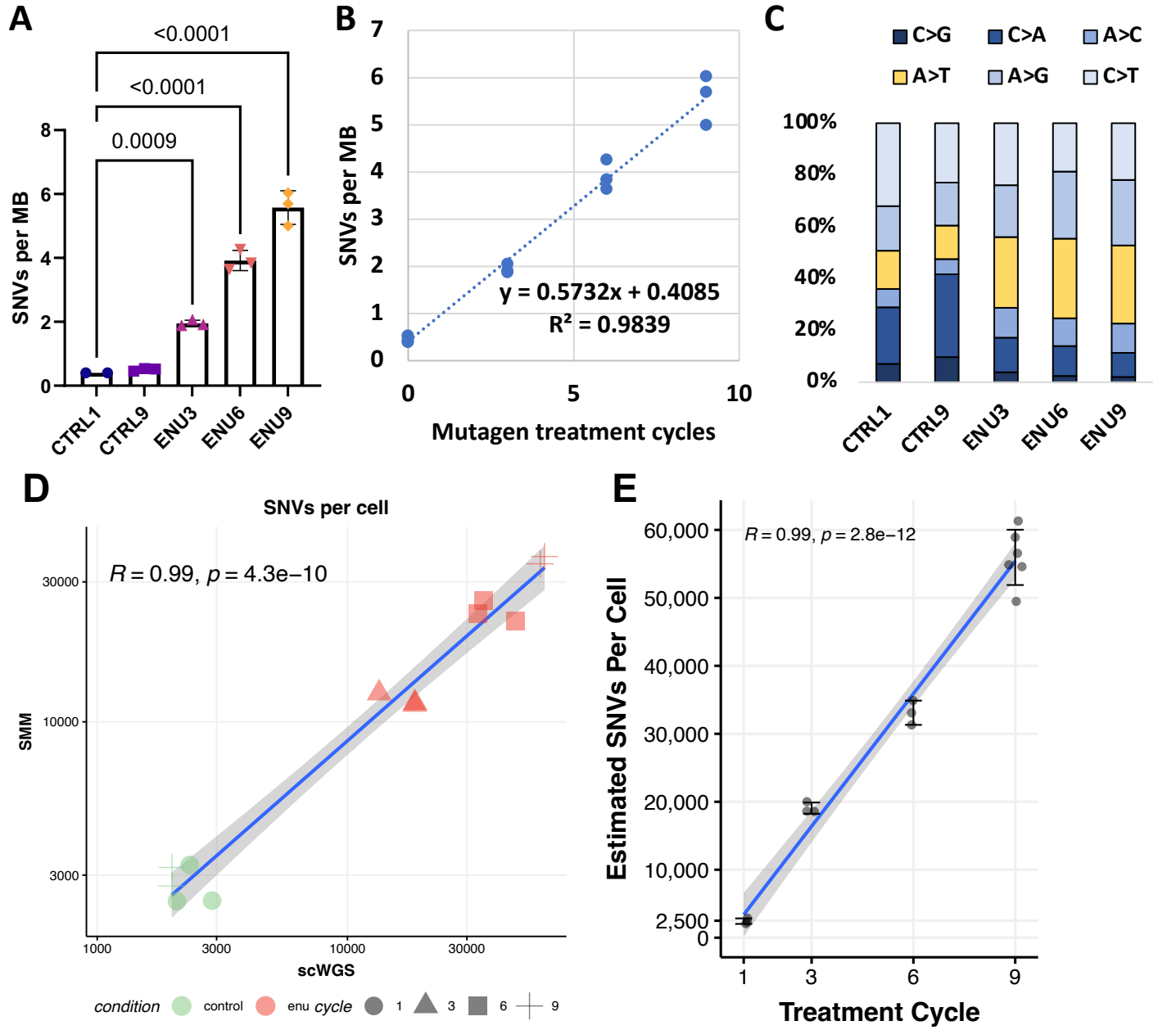
Supplementary Figure 1. Population doubling and cell death for cells collected after early and late recovery period (3 and 7 days post-treatment). A) Repeated treatment with low-dose ENU decreases population doubling (PD) rate over the course of the experiment when measured after 3 days and B) after 7 days (2-way ANOVA was used for significance). C) The relative PD time was significantly increased in ENU-treated cells compared to controls at the 3-day time point D) but not at the 7-day time point after treatment (significance was assessed with a t-test). E-H) Assessment of early (E-F) and late (G-H) apoptosis over time; with F and G) depicting the average percentage of apoptotic cells. (N=3, mean and standard deviation; paired t-test for determining significance of bar graphs; **: p<= 0.01).



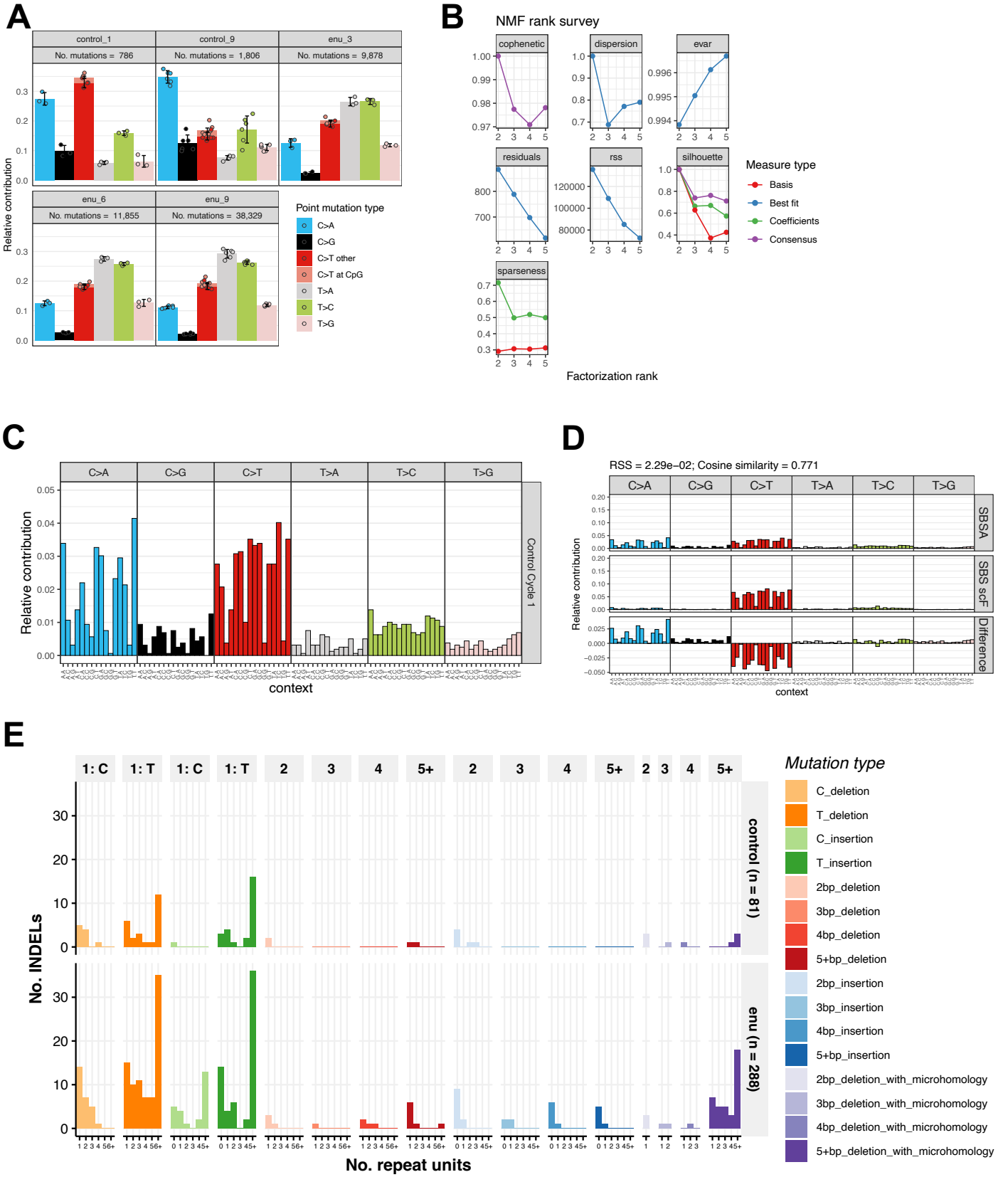
Supplementary Figure 2. Extended gene expression analysis of repeated mutagen treatment. A) Principal component analysis of gene expression data showed clear separation of control samples from cycle one (green) from control (orange) and ENU-treated (purple) samples from cycle nine. B) Heatmap of highly expressed genes revealed clustering of samples from experiment A and B apart from experiment C. C) MA-plots depicting large numbers of differentially expressed genes (DEGs) when comparing cells from cycle one to cycle nine (top), while control and ENU-treated samples from cycle nine result in only few DEGs across the three different experiments.



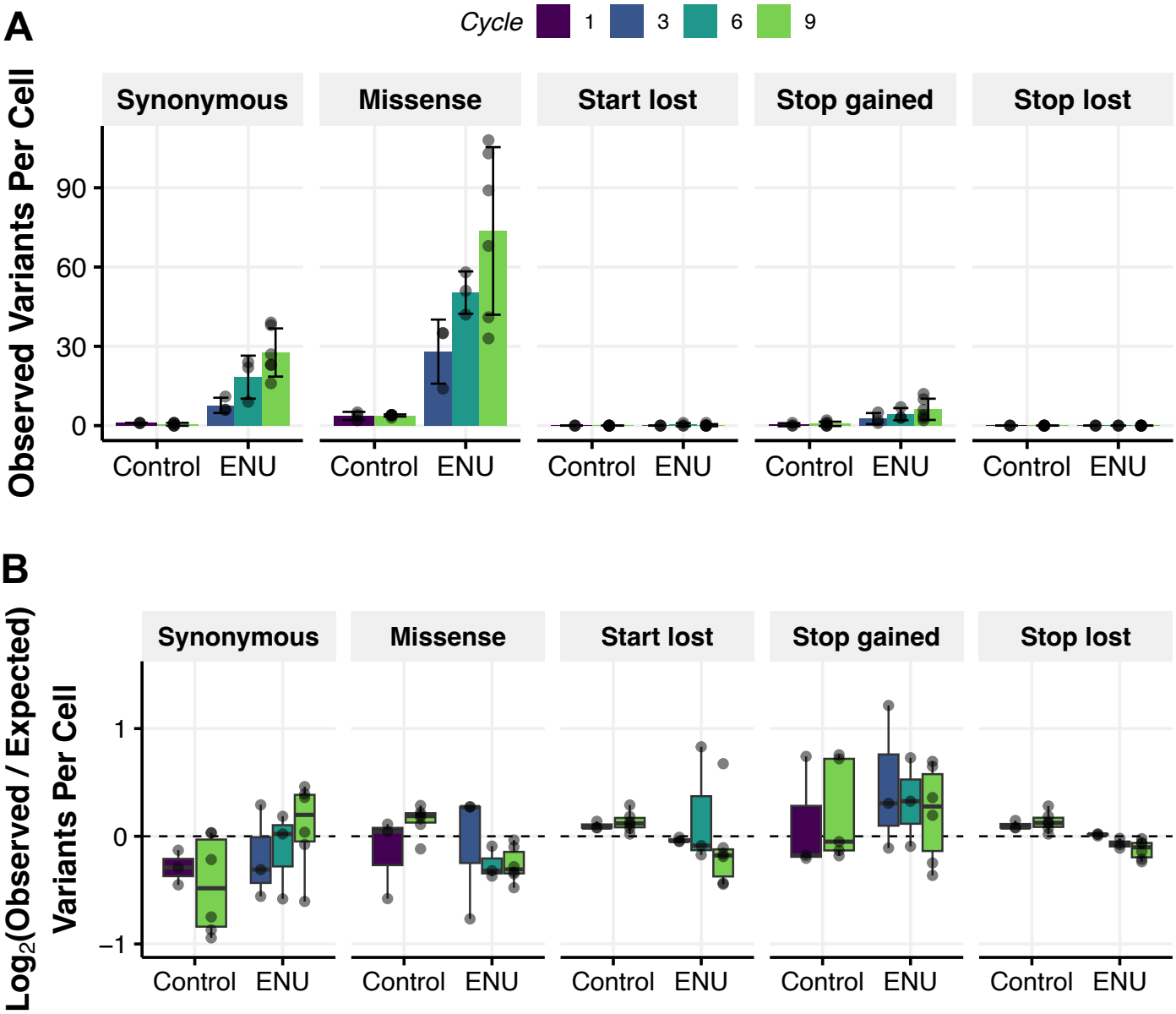
Supplementary Figure 3. Single-cell whole genome analysis quality control. A) Percent of genome covered at 20X. B-C) SNV and INDEL sensitivity. D-E) Number of SNVs and INDELs removed from samples collected in the same cycle and batch. F-G) Number of observed shared SNVs and INDELs after filtering for removal of shared mutations. H-J) Observed SNVs, INDELs, and INDEL/SNV ratio



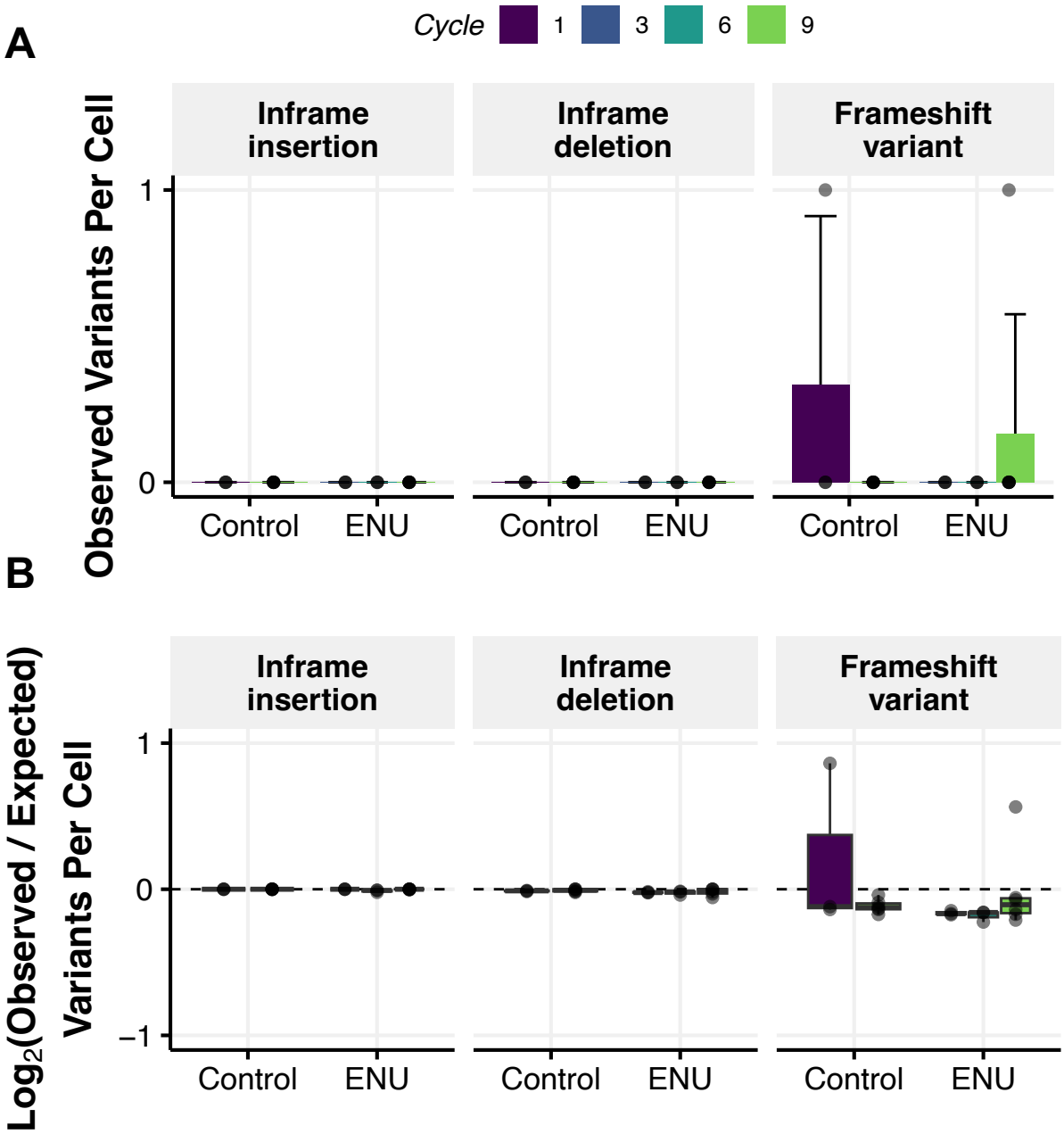
Supplementary Figure 4. Confirmation of somatic mutation burden using single-molecule mutation sequencing. A) Repeated treatment with low-dose N-ethyl-N-nitrosourea (ENU) increased somatic mutation burden at each cycle measured throughout the experiment using single-molecule assay. B) Single nucleotide variant (SNV) frequency increases linearly with the number of mutagen treatment cycles. C) Mutational spectrum of somatic mutations showing an increase in the proportion of T>A base transversions due to ENU induced mutations. D) Comparison of SNV burden between the single-cell whole genome sequencing or single-molecule mutation sequencing assays shows high correspondence. E) SNV frequency increases linearly with the number of mutagen treatment cycles using single-cell whole genome assay



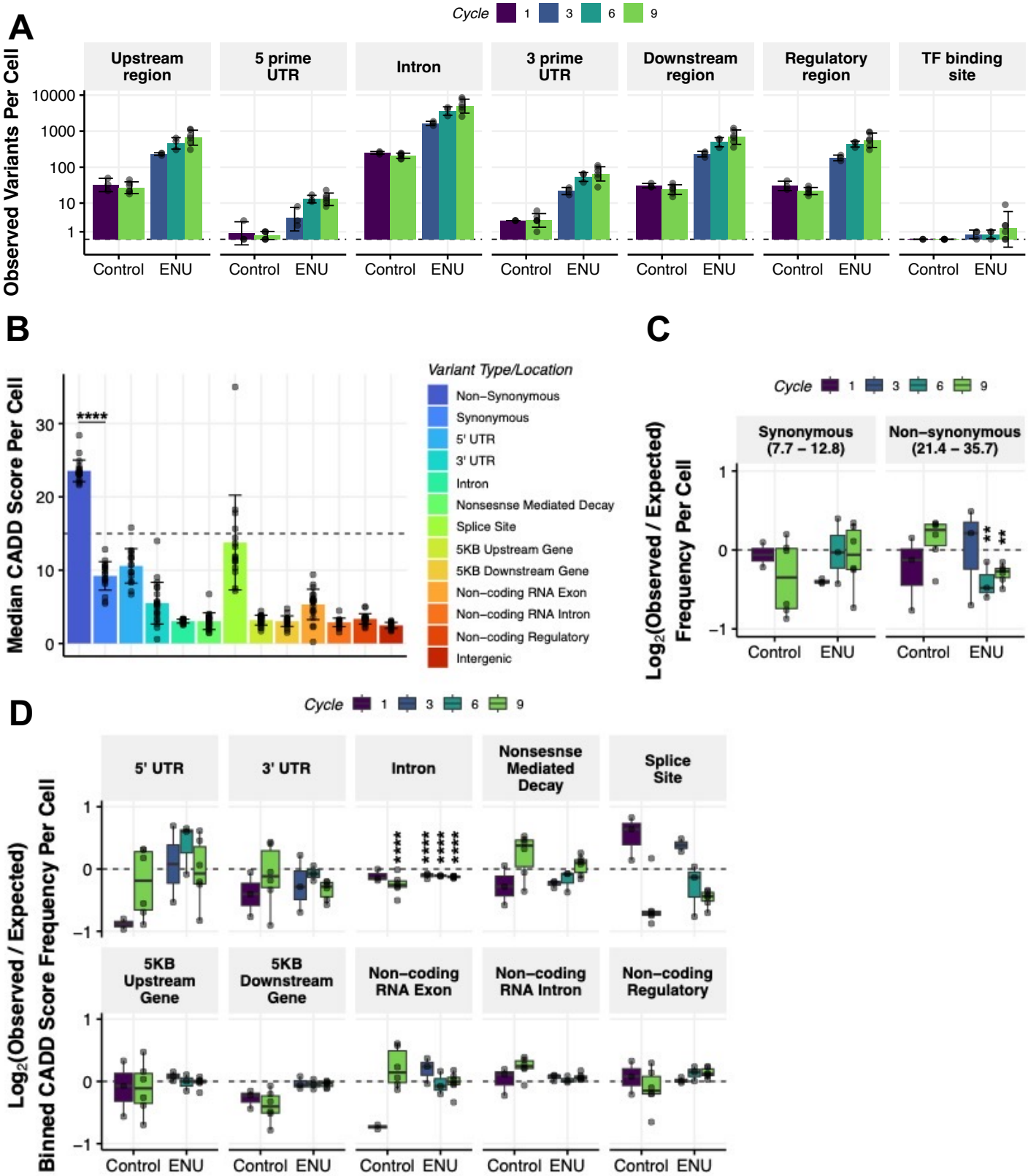
Supplementary Figure 5. Extended mutational signature analysis of SNVs and INDELs. A) Mutational spectrum separated by condition and cycle, showing abnormally increase C>T transition indicative of a technical artefact in control cycle 1 group. B) Results of determining optimal factorization rank for SNV signature extraction using non-negative matrix factorization where a rank of 2 was chosen. C) Mutational profile of control cycle 1 group. D) Comparing the 1 mutational signature, SBSA, extracted from the control cycle 1 cells compared to a known single cell library preparation artifact SBS scF. E) INDEL spectra for each condition.



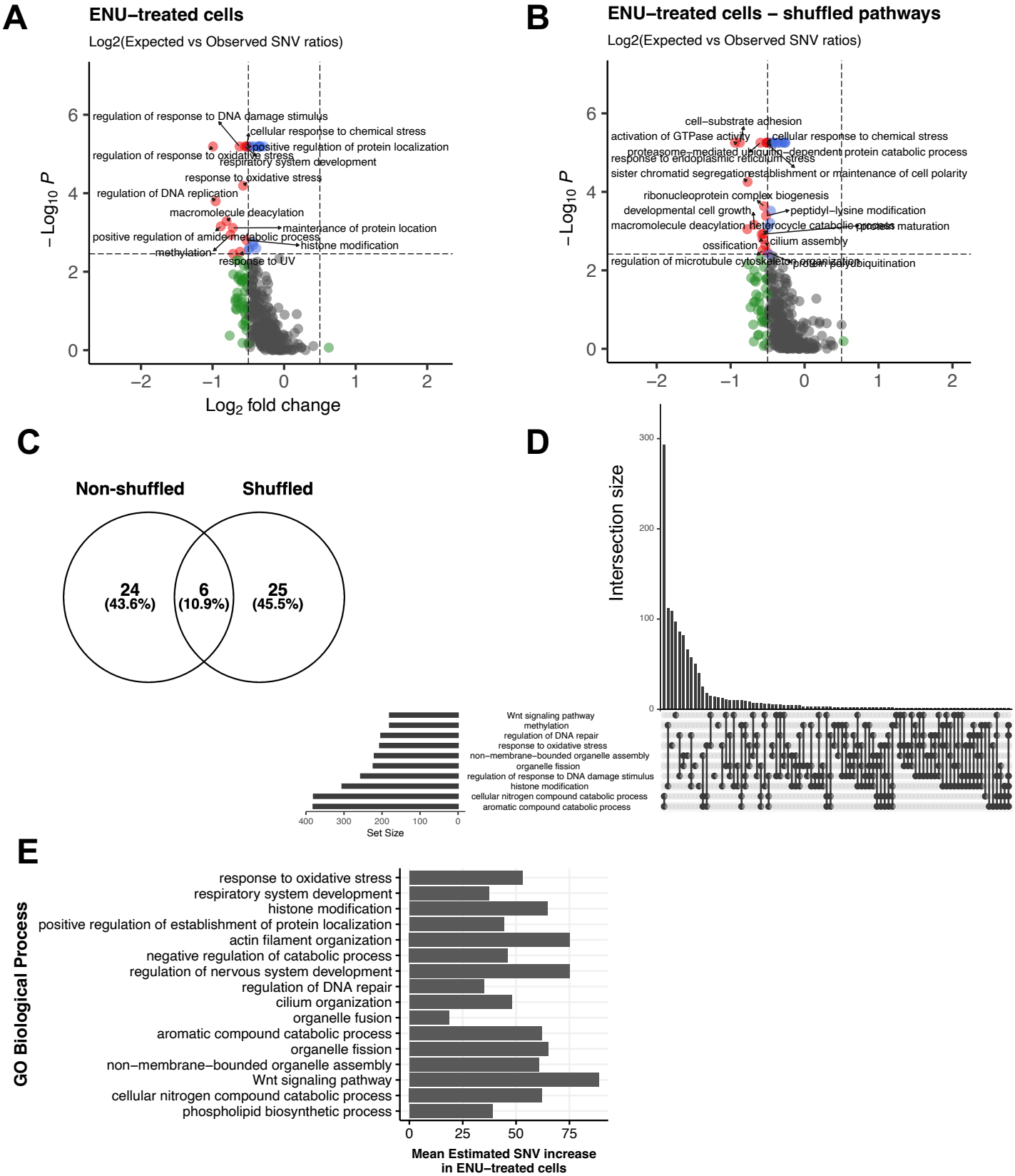
Supplementary Figure 6. Extended Analysis of Frequency and Selection Pressure of Damaging SNV Coding Mutations. A) Somatic mutation burden of specific SNV coding variants. Symbols at top of plots A-B indicate resulting p-values of Mann-Whitney test comparing control and ENU-treated groups with multiple hypothesis correction. B) Log_2 observed vs expected ratios of specific SNV coding variants. Symbols at top of plots C-D indicate resulting p-values of permutation test (N = 1000) comparing observed to expected; *: p <= 0.05; ***: p <= 0.001; ****: p <= 0.0001



Supplementary Figure 7. Extended Analysis of Frequency and Selection Pressure of Damaging INDEL Coding Variants. A) Somatic mutation burden of specific INDEL coding variants. Symbols at top of plots A-B indicate resulting p-values of Mann-Whitney test comparing control and ENU-treated groups with multiple hypothesis correction. B) Log_2 observed vs expected ratios of specific INDEL coding variants. Symbols at top of plots C-D indicate resulting p-values of permutation test (N = 1000) comparing observed to expected; *: p <= 0.05; ***: p <= 0.001; ****: p <= 0.0001



Supplementary Figure 8. Extended Analysis of Frequency and Selection Pressure of Non-Coding Variants. A) Somatic mutation burden of specific non-coding variants. B) Median scaled CADD score of all non-coding variants per cell grouped into specific subtypes based on genomic location. Symbols at top of plots A-B indicate resulting p-values of Mann-Whitney test comparing control and ENU-treated groups with multiple hypothesis correction. C) Log_2 observed vs. expected ratios of coding variants subdividing by their CADD score which corresponds to the median CADD score shown in B and reproduces the result shown in main figure 3C. D) Summary of Log_2 observed vs. expected ratios of binned CADD scores. Note that all variants, regardless of CADD score are shown as no significant results were found when subdividing non-coding variants by CADD score bins. Symbols at top of plots C-D indicate resulting p-values of permutation test (N = 1000) comparing observed to expected; *: p <= 0.05; **: p <= 0.01; ***: p <= 0.001; ****: p <= 0.0001



Supplementary Figure 9. Mutation distribution amongst gene ontology biological processes. A) Log_2 observed over expected ratios of mutations showing enrichment (>0) or depletion (<0) of mutations occurring in exons, promoters, and enhancers of genes apart of a biological process. B) Same as (A) except genes were randomly shuffled amongst the pathways to serve as a negative control. C) Overlap of pathways from Non-shuffled (A) and shuffled (C) results where mutations are significantly depleted (adjusted p-value < 0.05) that were then filtered out from the Non-shuffled result. D) Upset plot of top 10 pathways with overlapping genes from Non-shuffled (A) results where mutations are significantly depleted (adjusted p-value < 0.05). As expected similar pathways, such as cellular nitrogen and aromatic compound metabolic processes have many shared genes, will most other pathways do not which indicates diverse gene sets. E) Increase in mean estimated SNVs for pathways shown in Fig. 5C in the same order.